Exploring Strava's Suffer Score

Saryu Onishi*

April, 2023

Abstract

Training load quantification is a method of assessing the physiological impacts of a given workout. This kind of metric can provide coaches and athlete with valuable insights to how an athlete is responding to a prescribed workout. However, it is difficult to integrate into the decision making process as ease of access can be an issue. This project aims to estimate Strava's suffer score (a metric that quantifies training load) by using widely available activity summary metrics and the multiple linear regression model. To construct a viable multiple linear regression model, the backwards step-wise method of variable selection was used. This project found that by utilizing quadratic data transformation and interaction terms, a linear model would be able to estimate suffer score moderately well (Adjusted R-squared = 0.8). Hence this method of suffer score estimation could be suggested to individuals who do not have access to this metric, but are curious to see what they are.

*Department of Health and Exercise Sciences, University of Oklahoma. E-mail address: saryu@ou.edu

1 Introduction

Strava is a popular social media platform many athletes use to share their most recent athletic endeavours. These can range from walks around the block to snowboarding down a mountain. Even activities like pickleball can be uploaded onto Strava. In addition to the sharing, liking, and commenting on fellow users' uploads, athletes can use Strava to analyze their workouts. This is why Strava is particularly popular amongst endurance athletes. The analysis goes beyond the typical activity summaries GPS watches provide at the end of a workout. Although many of the workout analysis features can be accessed on the free version of Strava, some other features require users to be a paid member.

Strava's "suffer score" one of these feature that is only accessible to paid members. The suffer score is built to be a performance metric that represents how tough an activity was, in a single number. In exercise physiology, metrics like these are called measures of training load Bourdon et al. (2017). According to Strava (2016), their suffer score is calculated based on time spent in various heart rate (HR) 'zones'. The time spent in a higher HR zone is weighed heavier, compared to the time spent in a lower HR zone. Despite this measure being re-branded to "Relative Effort" (RE) recently, it will be referred to as suffer score in this project as this name remains in their data sets.

This project aims to estimate the Strava suffer score training load metric using simpler activity summary metrics to improve access, whether that may be for new users who are curious or athletes who do not have access to this feature/platform.

2 Literature Review

The term ”training load” is often used in exercise physiology to describe a type of metric that estimates the ”dose” of an exercise. This is because exercise physiologists think about exercise and their adaptations as a ”dose-response” relationship, where exercise provides a dose of physiological stress that our body responds to, through training adaptations (Lambert and Borresen (2010)). In a clinical commentary published by Paquette et al. (2020), they describe this metric in great detail. The authors explain that training load is defined as a product of external and internal load. External load is the physical and mechanical stress an exercise places on the body. These stresses are typically measures of time, distance or speed but can include others like ground reaction forces. Internal load is the physiological stress response that occurs due to the external load. Examples of internal load include measures of heart rate and perceived exertion (RPE). By calculating the product of these two types of measures, the training load takes both mechanical and physiological stresses applied to the body. In a way, this calculation weighs the external load to reflect the intensity of the activity.

In a review article, Lambert and Borresen (2010) discuss the concept of quantifying training load, as well as the various methods used to quantify training load. In this, the authors highlight the fact that there are many ways training load can be quantified, and there is no ”gold standard” that has been formally established. Different methods have various advantageous and disadvantageous. For example, lab equipment can be used to measure internal load through oxygen consumption and blood lactate data. However, these methods can be highly impractical for everyday use due to the cost and portability of the equipment. Other methods discussed include TRIMP, modified TRIMP and session RPE. The author writes that although all of these methods have strengths and

weaknesses, some are more desirable over others. The method that is most often used among the this list is TRIMP. TRIMP is an abbreviation of training impulse, and is a method of quantifying training load developed in the early 1990s as light weight and reliable heart rate monitors started to become available comercially.

Although the quantification of training load is accepted as an essential part of optimising training programs, the practical relevance of these measures are often under discussion amongst practitioners and athletes. In an article by Roos et al. (2013), the authors conducted several focus groups with elite coaches to identify what kinds of concerns existed amongst coaches, in terms of monitoring training through quantification of training load. In this, Roos et al. found that many coaches agreed that monitoring an athlete's physical condition in response to training was an important aspect of developing effective training plans. However, much of the challenge was in 1) obtaining good quality data from athletes at a consistent rate and 2) being able to respond to this data at promptly. To overcome these challenges, the coaches in the focus groups discussed the importance of feasibility of the monitoring system. In other words, distrust in the measure was not the issue. Instead, the concern was the ease of use when it came to utilising training load monitoring.

Strava's suffer score can be seen as one way to make measures of training load more accessible. Although this is only one way to quantify training load, it appears to use techniques similar to the TRIMPS method, using heart rate and duration as two of the main components Strava (2016). Although the exact formula is not published, the outlined method is in line with the current standards of training load quantification in exercise physiology. Hence, this could be considered as a metric that coaches and athletes can integrate into their repertoire, when assessing physiological impact of any given workout.

3 Data

Data used in this project comes from Strava activity logs. Because Strava is a social platform where activity data is shared, it retains a record of all uploaded activities. The data is mostly numerical and consist of summary metrics of the activities. This includes duration, distance, speed, elevation gain and loss. If applicable, heart rate averages and maximums, cadence averages and maximums, and power averages and maximums are also present. These depend on whether the user utilizes technology such as heart rate monitors and power meters.

In addition to the activity summary metrics, the data set includes location data, activity descriptions, social interaction metrics (kudos and comment count) and foreign keys for gear selection (if recorded), maps, and the activity stream. The activity stream data table allows Strava to store more data on a single activity, enabling the in-depth analysis features. This a separate data table is tied to the activity log data through the activity stream foreign key (activity ID).

3.1 Data Collection

Data for this project was scraped from the activity log of a single Strava user, using the Strava API and the rStrava R package. The data was then loaded into R studio software for further analysis. There are a few reason why data was only collected from one user. Firstly is the authentication requirements that comes with scraping activity log from Strava users. Each request for data prompts the target user to authenticate the access. Hence, it was most practical to use a single user profile that was easily accessible for this project. In addition to the authentication requirements, physiological responses to exercise can drastically differ between individuals and environmental conditions. Because physiological responses to exercise will be the main predictors

of suffer score (Strava's training load measure), it was decided that collecting data from other users was not imperative to meet the goals of this specific project.

4 Methods

The primary method of analysis used was the backwards step-wise multiple regression, with the goal of analyzing the magnitude of impact of relevant metrics on the Suffer Score. The multiple regression model can be depicted by the following equation:

$$Y_r = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \varepsilon, \tag{1}$$

For this project, Y_r represents the Strava Suffer Score for recorded runs (r), and X_n are summary metrics. The parameters of interest were the β coefficients, as this shows the degree of influence each metric has towards estimating the Suffer Score.

4.1 Data Cleaning

Before the analysis, the activity log was filtered to only include summary metrics such as time, distance, average heart rate and average speed. This was done because the goal of the project was to determine what summary metrics could be used make the best estimation the Suffer Score. The data was also filtered to only include runs, as physiological responses to cycling and running (the two main types of activities in this data set) can differ significantly Hassmén (1990). Finally, the data was filtered to only include runs that had heart rate data available. This was because heart rate was determined to be a key component of training load quantification, and Strava would not have

generated a Suffer Score in the absence of heart rate data.

The data was then mutated. First, some of the data had to be converted to the numeric class. This was an important step to ensure compatibility between the data and the analysis method. Additionally, the time variable was converted from time in seconds, to time in minutes. This was done to make the data easily understandable and user friendly.

4.2 Data Analysis

For the analysis, a backwards step-wise multiple regression was conducted. The general method of this started with creating a linear model with all the available variables. This initial model is then put through the process of variable selection where the statistically non-significant variable is removed. In R, this was done using the lm() function for the initial regression, and the step() or stepAIC() functions for the step-wise regression.

In the project, three main models were constructed, each with increasing complexity. The first model was straightforward backwards step-wise multiple regression in which no variables were manipulated. The next model was an extension of the first model, with added quadratic terms. The final model was an extension of the second model, with the addition of interaction terms between a time or distance and heart rate or cadence.

5 Research Findings

The data used in the models are summarized in table 1. There were 306 observations in the final data set and the mean suffer score ranged between 2 and 208, with a mean of 47.7. The models were summarized in table 2, in order of complexity. 'model 1' is the standard backwards step wise

multiple regression, 'model 2' includes quadratic terms, and 'model 3' includes quadratic terms and interactions.

The most significant predictors of suffer score appears to exclude elevation gain in all three models. For the first model, the only predictors included in the final model are distance, time and average heart rate. All of these are statistically significant ($p < 0.05$). Model 2 excludes distance, but includes maximum heart, in addition to the quadratic terms of distance, time, speed and heart rate ($p < 0.05$). Model 3 includes 13 predictors. 6 of these are simple summary metrics, 3 are quadratic terms, and the remaining 4 are interaction terms between distance/time and heart rate/cadence.

The summary of the three models (Table 2) show that the most complex version that included quadratic terms and interactions was able to estimate the outcome variable Suffer Score to a greater level of accuracy (adjusted R-squared of 0.8 and RMSE of 14.65). With that said, the least accurate model also performs fairly well, in terms of R-squared and RMSE values (0.7 and 18 respectively).

When the predictions of the models were plotted against the actual suffer score values, these performance differences can be put into context. The performance of the first model is visualized in figure 1, where the x axis contains predicted suffer score values and the y axis contains the actual suffer score values. The prediction accuracy gets increasingly poor as the suffer score values increase. This indicates high heteroscedasticity, or varying error depending on the value. Furthermore, many negative suffer scores are predicted, which are invalid predictions.

Figure 2 shows the performance of model 2. Relative to model 1, model 2 has achieved some improvements, particularly, at higher values. This is signified by the tighter distribution of points around the blue line that represent perfect fit. However, this model also predicts many negative suffer scores.

The adjusted R-squared values and RMSE values indicate that model 3 has the best performance. This is visually demonstrated in figure 3. Predictions are fairly consistent from low to high values. Although the negative suffer score predictions are still present, it is to a lesser degree, compared to the previous models.

6 Conclusion

The analysis of summary metrics to predict suffer score has demonstrated the limits of utilizing simple activity summary metrics and linear regression methods to accurately estimate suffer score. This does not come as a surprise because it is likely that Strava utilizes a more complex calculation technique, or integrates more in-depth data from activity streams to calculate their suffer score. With that said, the model predictions were not too far of the actual suffer score. The analysis showed that when using summary metrics, quadratic transformations, and interaction terms between some of the summary metrics, a linear model can account for approximately 80% of the variation.

Ultimately, Strava's suffer score is only one version of training load quantification available to athletes. Additionally this metric is highly individual and open to interpretation. There are no clear ties between our physiological condition and the suffer score, that has been shown in scientific literature. Hence, for an everyday, non-premium Strava user, this kind of model may be sufficient, as it could satisfy their curiosities of the suffer score metric.

For future research, the model developed in this project could be tested with data from other users to measure the degree of individuality of the suffer score metric. Additionally, machine learning and the activity stream data could be integrated to generate a model that may estimate

Suffer Score better. However, more interesting research opportunities may come as a form of a validation study, comparing this and other training load quantification to subjective measures like fatigue and rate of perceived exertion, and physiological markers of stress.

References

Bourdon, Pitre C, Marco Cardinale, Andrew Murray, Paul Gastin, Michael Kellmann, Matthew C Varley, Tim J Gabbett, Aaron J Coutts, Darren J Burgess, Warren Gregson et al. 2017. "Monitoring athlete training loads: consensus statement." *International journal of sports physiology and performance* 12 (s2):S2–161.

Hassmén, Peter. 1990. "Perceptual and physiological responses to cycling and running in groups of trained and untrained subjects." *European journal of applied physiology and occupational physiology* 60:445–451.

Lambert, Michael Ian and Jill Borresen. 2010. "Measuring training load in sports." *International journal of sports physiology and performance* 5 (3):406–411.

Paquette, Max R., Christopher Napier, Richard W. Willy, and Trent Stellingwerff. 2020. "Moving Beyond Weekly "Distance": Optimizing Quantification of Training Load in Runners." *Journal of Orthopaedic & Sports Physical Therapy* 50 (10):564–569.

Roos, Lilian, Wolfgang Taube, Monika Brandt, Louis Heyer, and Thomas Wyss. 2013. "Monitoring of daily training load and training load responses in endurance sports: what do coaches want?" *Schweizerische Zeitschrift für Sportmedizin und Sporttraumatologie* 61 (4):30–36.

Strava. 2016. "Suffer Score: How Hard is Hard?" `https://blog.strava.com/suffer-score-how-hard-is-hard-11775/`. Accessed: 2023-04-24.

Figures and Tables

Table 1: Statistics for Strava Run Data

Metric (units)	Unique (#)	Mean	SD	Min	Median	Max
suffer_score	110	47.7	33.5	2.0	42.0	208.0
distance	341	11.4	4.6	0.9	11.0	23.5
time_minutes	335	59.1	24.5	4.5	58.0	44.7
average_speed	306	12.4	1.6	8.4	12.4	20.2
total_elevation_gain	281	54.9	59.7	0.0	36.2	303.1
average_heartrate	239	147.6	11.8	116.6	147.2	184.0
max_heartrate	64	169.5	13.4	130.0	172.0	204.0
average_cadence	83	88.1	2.2	76.8	87.8	97.5

Notes: Sample size for all variables is $N = 306$.

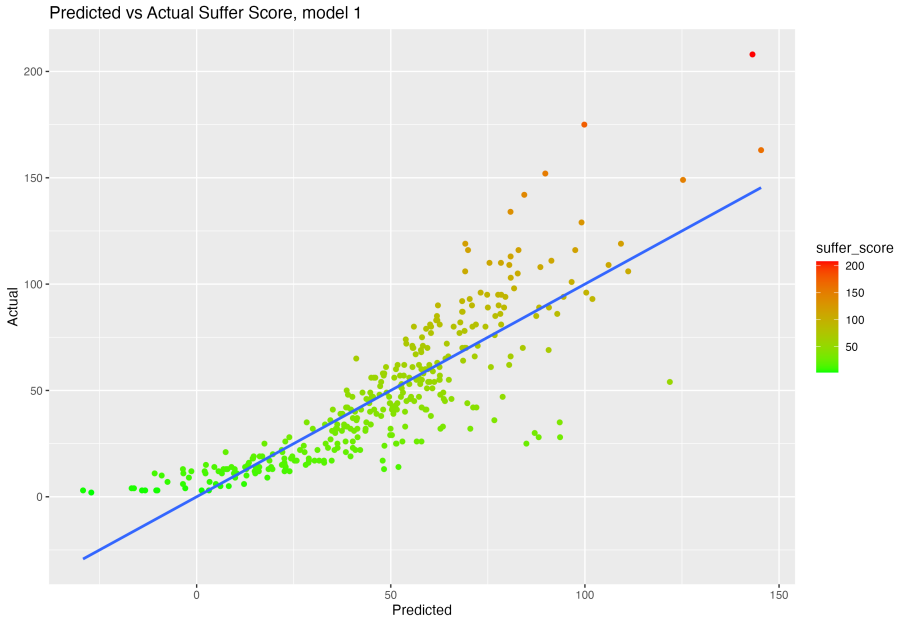

Figure 1: Predictions of Model 1

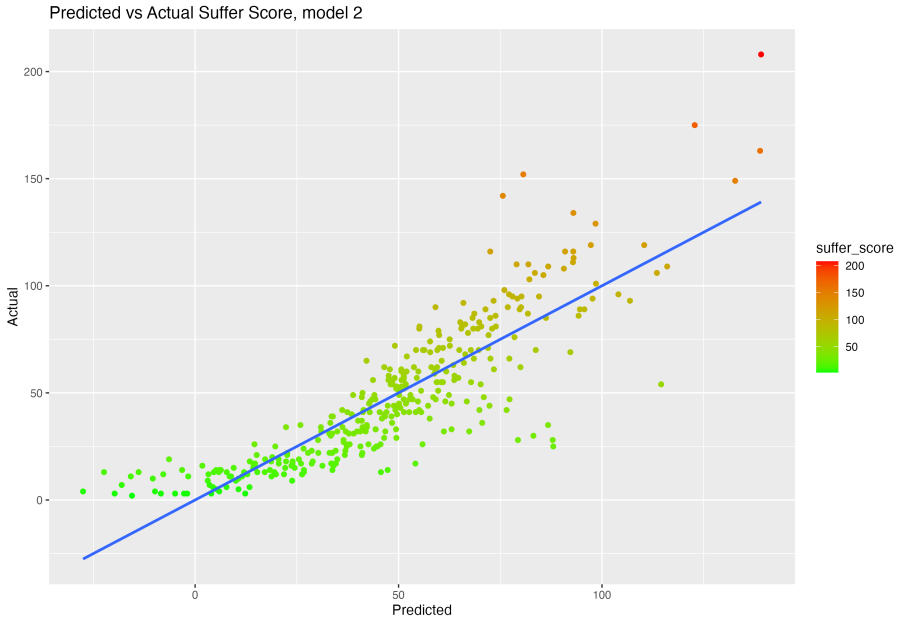

Figure 2: Predictions of Model 2

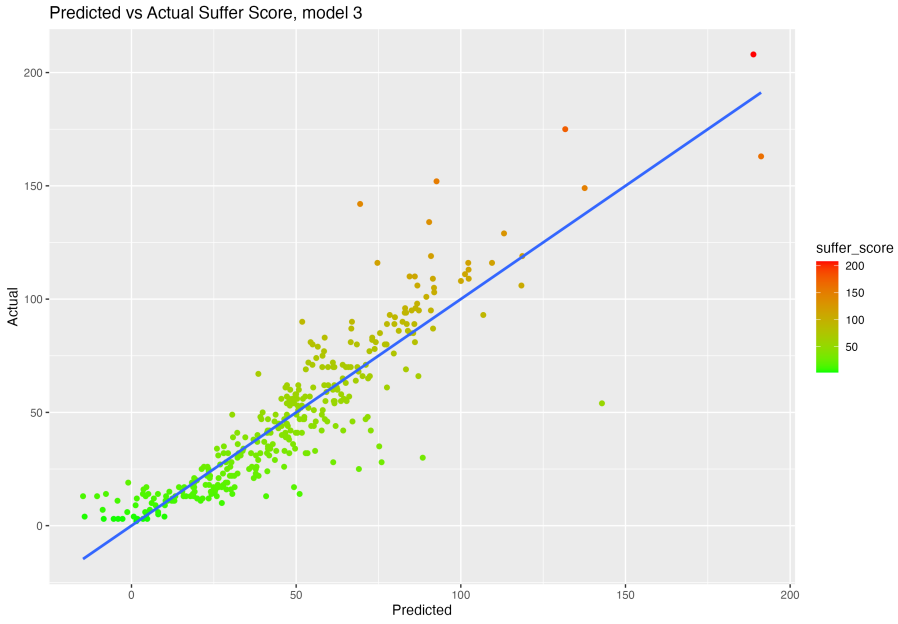

Figure 3: Predictions of Model 3

Table 2: Model Summaries

	model 1	model 2	model 3
(Intercept)	-213.131***	-177.770	158.134
	(¡0.001)	(0.153)	(0.291)
distance	7.495***		7.951
	(¡0.001)		(0.732)
time_minutes	-0.599***	0.711***	-9.337*
	(¡0.001)	(¡0.001)	(0.042)
average_heartrate	1.430***	9.855***	11.070***
	(¡0.001)	(¡0.001)	(¡0.001)
max_heartrate		-8.051***	-6.948***
		(¡0.001)	(¡0.001)
I($distance^2$)		0.205***	
		(¡0.001)	
I($time\_minutes^2$)		-0.007***	
		(¡0.001)	
I($average\_speed^2$)		0.053	0.802**
		(0.107)	(0.003)
I($average\_heartrate^2$)		-0.028***	-0.035***
		(¡0.001)	(¡0.001)
I($max\_heartrate^2$)		0.024***	0.021***
		(¡0.001)	(¡0.001)
average_speed			-18.419*
			(0.011)
average_cadence			-3.986***
			(¡0.001)
distance × average_heartrate			0.440***
			(¡0.001)
distance × average_cadence			-0.771*
			(0.016)
time_minutes × average_heartrate			-0.066***
			(¡0.001)
time_minutes × average_cadence			0.214***
			(¡0.001)
R2 Adj.	0.708	0.732	0.801
RMSE	18.00	17.11	14.65

Notes:+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$